Final Report

COMPUTERIZED ABILITY TESTING
1972 - 1975

David J. Weiss

April 1976

PSYCHOMETRIC METHODS PROGRAM
DEPARTMENT OF PSYCHOLOGY
UNIVERSITY OF MINNESOTA
MINNEAPOLIS, MN 55455

FINAL REPORT OF PROJECT NR150-343, N00014-67-A-0113-0029
SUPPORTED BY THE
PERSONNEL AND TRAINING RESEARCH PROGRAMS
PSYCHOLOGICAL SCIENCES DIVISION
OFFICE OF NAVAL RESEARCH

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
Final Report: Computerized Ability Testing, 1972-1975

5. TYPE OF REPORT & PERIOD COVERED
Final Report.
March 1972-September 1975

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
David J. Weiss

8. CONTRACT OR GRANT NUMBER(s)
N00014-67-A-0113-0029

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Department of Psychology
University of Minnesota
Minneapolis, Minnesota 55455

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
P.E.:61153N PROJ.:RR042-04
T.A.:RR042-04-01
W.U.:NR150-343

11. CONTROLLING OFFICE NAME AND ADDRESS
Personnel and Training Research Programs
Office of Naval Research
Arlington, Virginia 22217

12. REPORT DATE
April 1976

13. NUMBER OF PAGES
22

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)
Approved for public release; distribution unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

testing	sequential testing	programmed testing
ability testing	branched testing	response-contingent testing
computerized testing	individualized testing	automated testing
adaptive testing	tailored testing	

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Three and one-half years of research on computerized ability testing are summarized. The research program's original research objectives are described, and the research approach is summarized and related to the eighteen Technical Reports produced under this contract. Twenty-one major research findings are presented. The implications of the research findings and methods for future research in computerized adaptive testing are described. Also included are abstracts of the eighteen Technical Reports derived from this research.

CONTENTS

FINAL REPORT:
COMPUTERIZED ABILITY TESTING, 1972-1975

Objectives

The original objectives of the research were:

1. To develop and implement the stratified computer-based ability test.

2. To compare, on psychometric criteria, the various approaches to computer-based ability testing, including:

   a. The stratified computerized test
   b. The pyramidal approach
   c. Lord's flexilevel test
   d. Two-stage testing
   e. Some of the mathematical models for computerized testing.

3. To determine the effect on ability test scores of:

   a. Maintaining test items at a level of difficulty near the individual's estimated ability level by means of computer-controlled administration under one or more of the above strategies.

   b. Providing "feedback" of correctness of response to each item within the standard models for computerized testing by using a special variation of the stratified approach designed to insure various proportions of correct responses.

4. To determine the utility for diagnostic purposes of information on an individual's item response latencies.

Research in pursuance of these objectives began in March 1972 and continued through September 15, 1975. The research led to the publication of sixteen Technical Reports, with two more currently in preparation. Abstracts of all eighteen Technical Reports follow this overview of the research program.

Approach

Research began with a comprehensive review of the literature on adaptive or tailored testing (Research Report 73-1). This review identified four major research approaches to problems of adaptive testing. These included empirical (live-testing) studies, monte carlo computer simulation studies, "real-data" simulation studies, and theoretical studies. These research approaches were evaluated, and it was concluded that live-testing studies and monte carlo simulation studies provided the most useful kinds of research information. This review of the literature also led to the conclusion that very little was known about the

various strategies of adaptive testing. The little evidence that was available suggested that adaptive tests provided better measurement than conventional tests under a variety of circumstances. But it was obvious that considerable research was necessary in order to discover appropriate measurement applications, approaches, and theory relevant to adaptive testing.

The review of the literature identified a number of different strategies of adapting ability tests to individuals. These strategies were described in Research Report 74-5. Each strategy was evaluated in terms of its potential for providing equiprecise measurement (measurements with equal precision at all levels of ability), and in terms of its feasibility under both paper-and-pencil and computer administration.

In order to provide an item pool for live-testing studies, a pool of 575 multiple-choice word knowledge items was calibrated (Research Report 74-2). From this pool, 369 items were used in all subsequent live-testing studies implemented in the research program. Word knowledge items were selected because of their general use in "intelligence" tests and their appearance in almost all major multiple-aptitude batteries. Development of the item pool and its refinement led to an analysis of its dimensionality, and a general purpose computer program developed to assist in that analysis was described in Research Report 75-2.

The development of the stratified adaptive (stradaptive) computerized test was reported in detail in Research Report 73-3. Subsequent live-testing research with this adaptive testing strategy (Research Report 75-4) and computer simulation studies (Research Report 75-6) were implemented in order to evaluate its characteristics and feasibility.

The combination of live-testing and computer simulation studies was continued in the investigation of the psychometric characteristics of other strategies of adaptive testing. Live-testing research with the two-stage adaptive testing strategy was reported in Research Report 73-4, and computer simulation studies replicating and extending those findings are in Research Report 74-4. Research Report 75-3 presents the results of both live-testing and computer simulation studies of the flexilevel strategy. Computer simulation data using a Bayesian adaptive testing strategy, motivated by findings of live-testing research, are in Research Report 76-1. Data from live testing with the pyramidal adaptive testing strategy are reported in Research Report 74-3.

One study (Research Report 75-1) presents empirical data comparing two adaptive testing strategies. Although the original plans included a substantial number of these inter-strategy comparisons, it became evident that it is difficult to draw clear conclusions concerning the comparison of two or more strategies of testing from live-testing studies. Consequently, later attempts to compare the relative effectiveness of different strategies of adaptive and/or conventional testing utilized computer simulation studies. Research Report 75-5 presents the results of a computer simulation study comparing the psychometric

characteristics of a number of adaptive testing strategies and a discussion of some of the problems involved in live-testing studies.

Live-testing research on adaptive testing strategies began using cathode-ray tube terminals (CRTs) acoustically coupled to a time-shared computer system (Research Report 74-1). Because the characteristics of this computer system did not permit research in furtherance of Objectives 3 and 4, in 1974 the research program began using a minicomputer. This system allowed accurate measurements of testee response latencies and an environment permitting the study of the psychological effects of computerized testing.

Studies of the effects of "feedback" or knowledge of results, on ability test scores and testees' psychological reactions, are reported in Research Reports 76-3 and 76-4. A first analysis of testee item response latency data is reported in Research Report 76-2.

Major Findings

The major findings summarized below are generally organized according to the original objectives of the research program. Additional details are in the Research Report abstracts. Many of the original Research Reports contain additional important findings concerning specific adaptive testing strategies or methodological aspects of research in adaptive testing.

1. Implementation of the stratified adaptive (stradaptive) computerized test shows that it is a feasible approach to computerized adaptive testing (Research Report 73-3). Evaluation of the stradaptive test in comparison with other strategies of adaptive testing (Research Report 74-5) shows that it has considerable logical appeal as a result of its use of differential entry points, a flexible termination criterion which can take account of guessing, and efficient use of real item pools. The stradaptive test also provides scores which reflect the consistency with which a testee interacts with an item pool.

2. Simulation research comparing the stradaptive test with other strategies of adaptive ability measurement, under one item pool configuration, shows that its information curve is the flattest of the strategies studied (Vale, in Research Report 75-5). Thus, of the adaptive testing strategies studied, the stradaptive test appears to provide the best realization of the ideal of measurement with equal and high precision at all trait levels.

3. Research comparing the stradaptive test with non-adaptive approaches to ability testing (Research Report 75-6) shows that the stradaptive test provides more equiprecise measurement than a peaked conventional test. As item discriminations increased, the stradaptive test provided a greater advantage

in terms of equiprecision. While a rectangularly distributed conventional test also provides equiprecise measurement, the level of precision is substantially lower than that of an otherwise comparable stradaptive test (Vale, in Research Report 75-5).

4. Live-testing research with the stradaptive test (Research Report 75-4) shows that its consistency scores, which appear to reflect the dimensionality of the interaction of an individual with a given item pool, show promise of being good moderator variables for the prediction of test-retest stability. This research showed very high test-retest stabilities for a highly consistent group of individuals, in comparison to only moderate test-retest stabilities for those individuals whose consistency scores on first testing were lower.

5. Rational comparison of adaptive testing strategies (Research Report 74-5) suggested that some of the approaches proposed for adaptive testing (e.g., the Robbins-Monro procedure) are infeasible with real item pools. Bayesian and maximum likelihood approaches to adaptive testing appeared to be the most promising, followed by the stradaptive test and the pyramidal models. This evaluation also suggested that the flexilevel test had the least logical appeal of the adaptive testing models proposed.

6. In a computer simulation study (Vale, in Research Report 75-5), all of the adaptive testing strategies provided more equiprecise measurement than did a peaked conventional test. All of the adaptive strategies provided higher levels of average information than did a rectangular conventional test.

7. The computer simulation study (Vale, in Research Report 75-5) also provided comparative information on the relative equiprecision of adaptive testing strategies. Within the adaptive testing strategies, the rankings of the strategies based on the obtained information curves were about the same as those based on the previous logical evaluation of the strategies. Thus, the stradaptive and Bayesian strategies yielded the most desirable measurement characteristics and the flexilevel test provided the least desirable characteristics.

8. Computer simulation and live-testing studies (Research Report 75-3) indicated that the flexilevel test offered little improvement in measurement characteristics over a conventional peaked test. In addition, the flexilevel test was evaluated as having the potential to raise negative psychological effects as a result of its branching strategy (Research Report 74-5).

9. In live-testing studies comparing the test-retest stabilities of adaptive testing strategies and peaked conventional tests (Research Reports 73-4, 74-3, 75-3, 75-4) when tests were equated for number of items and memory effects, the adaptive tests generally had higher test-retest stabilities than did the conventional test. There were no major differences between the test-retest stabilities resulting from the different adaptive testing strategies.

10. While a Bayesian adaptive testing strategy was logically evaluated as a promising testing strategy (Research Report 74-5) and yielded information curves in one study which had desirable characteristics (Vale, in Research Report 75-5), research using different criteria of evaluation showed that this testing strategy has some problems which reduce its utility (McBride, in Research Report 75-5; Research Report 76-1). Both live-testing and computer simulation studies showed that the Bayesian adaptive testing strategy studied yielded scores which were highly correlated with test length. In addition, ability estimates derived from this strategy were biased for two-thirds of the typical ability range. This testing strategy also yields scores which are dependent upon the characteristics of the prior ability estimate required by the testing strategy.

11. A major problem in the implementation of two-stage adaptive tests is that of misclassification due to errors of measurement in the routing test (Research Reports 73-4, 74-4). But a well-designed two-stage test can provide information curves which are flatter, hence yielding more equiprecise measurement, than a peaked conventional test (Research Report 74-4; Vale, in Research Report 75-5).

12. A simple pyramidal adaptive testing strategy (Research Reports 74-3, 75-1) with a fixed step size is a promising approach to adaptive testing. Its major problem is that it results in information curves which are low for ability levels divergent from the mean (Vale, in Research Report 75-5). But it appears to provide a wider range of adequate measurement than conventional tests, or the two-stage or flexilevel adaptive testing models. In terms of providing equiprecise measurement, however, its results are not as good as those of the stradaptive test or the Bayesian adaptive test.

13. Implementation of adaptive testing requires the use of scoring methods other than simple number-correct scores such as average difficulty of correct responses, or difficulty of last item answered. Live-testing research with several alternative scoring methods (Research Reports 74-3, 75-4) shows that they provide scores with different characteristics in terms of test-retest stabilities, distributional characteristics, and correlations with other variables. Computer

simulation research (McBride, in Research Report 75-5) shows that scoring methods derived from latent trait theory differ in terms of bias, regression on ability, and precision. Further research is needed on the development of optimal scoring methods for adaptive testing.

14. The evaluation of competing strategies of testing, including competing strategies of adaptive testing, is quite difficult (Sympson, in Research Report 75-5). Live testing studies are complicated by memory effects, lack of adequate criteria, and non-compatability of tests due to differing test lengths and differing item discriminations. Live-testing studies comparing adaptive and conventional tests may also be complicated by psychological effects (Research Reports 76-3 and 76-4). Computer simulation studies can provide evaluations on a variety of criteria, but it is necessary that the computer simulation model be shown to adequately reflect the behavior of real testees (Research Reports 74-4, 75-3, 75-6). Theoretical or analytic studies are extremely limited as a means for evaluating adaptive testing. Because of the restrictive assumptions necessary to implement theoretical studies, they can provide only limited conclusions which may not generalize to more realistic conditions. It appears that the best approach to evaluating competing testing strategies is a systematic combination of live-testing studies and extensive computer simulations.

15. An analysis of response latency data shows that testees approach different testing procedures in different ways (Research Report 76-2). The response latency data suggest that these different test-taking styles and strategies might be potentially useful as moderator or predictor variables in the prediction of external criteria.

16. Computer-administered feedback (immediate knowlege of results) on a conventional test appears to result in enhanced ability test performance for testees of all ability levels (Research Report 76-3). Under computer-administered feedback conditions, mean test scores were significantly higher for both high- and low-ability testees. Ninety percent of college student testees favorably evaluated their experience with computer-administered feedback (Research Report 76-4).

17. Adaptive tests appear to be intrinsically more motivating for low-ability testees (Research Report 76-4), and result in higher ability estimates (Research Report 76-3), than similarly administered conventional tests. This suggests that adaptive testing might eliminate some of the undesirable psychological effects characteristic of conventional testing procedures, resulting in fairer and more accurate test scores for testees who typically obtain low scores on conventional ability tests.

18. In a computer-administered conventional test, the provision of immediate feedback (immediate knowledge of results) to minority group members appears to raise their ability scores to the levels of members of the non-minority group (Betz, in Research Report 75-5). This effect appears to be a motivational effect, similar to the motivational effect found for adaptive tests administered to other testees who typically obtain lower average scores on conventional ability tests (Research Reports 76-3 and 76-4).

19. Computerized adaptive testing results in tests of fewer items, to achieve the same or higher degrees of accuracy, than conventional tests. Consequently, with total testing time fixed, the use of computerized adaptive testing results in extra time available during testing which can be used in the measurement of other abilities or in obtaining measurements of higher accuracy. More precise measurement, or data available from the measurement of other abilities, combined with additional data available from computerized testing (e.g., response consistency, Research Report 75-4; response latency, Research Report 76-2), might result in increased validity in the prediction of external criteria.

20. The implementation of adaptive testing on large-scale time-shared computer systems can be hazardous. Experience in testing thousands of students on a system of this type shows that it is too unpredictable to provide a reliable means of adaptive testing. In addition to frequent failures due to hardware, software or communications problems, computer response time was generally too long and too variable. These characteristics combined to raise the possibility of negative psychological effects among testees who were being tested on the system. Experience with a dedicated minicomputer system in this research program indicates that this approach to adaptive testing is considerably more desirable than the use of a large-scale time-shared computer system, with little increase in cost.

21. A preliminary cost analysis, based on the administration of adaptive tests by minicomputer, suggests that computerized adaptive testing will be a financially feasible approach to ability testing. Given the continuing and projected decreases in costs of computer equipment, it is expected that the costs of adaptive testing will approximate those of paper-and-pencil test administration and scoring within a few years.

Implications for Further Research

The findings and experience of this three-and-one-half-year research program support the feasibility, utility and psychometric advantages of computerized adaptive ability testing. However, many new questions were

raised by the research, and some of the original questions addressed are still in need of further research. Portions of the research described below are being pursued under a contract entitled "Computer-Based Adaptive Measurement of Intellectual Capabilities", NR 150-382, with the Personnel and Training Research Programs of the Office of Naval Research.

Branching strategies. While the research program has answered some preliminary questions concerning the relative utility of various branching strategies for use in adaptive testing, considerably more research is needed. The evaluations done thus far have used both live-testing and computer simulation studies. The results of live-testing studies were confounded by memory effects, differing item discriminations, and by potential psychological effects.

Studies designed to equate testing strategies for both memory effects and differing item discriminations are currently in progress. Some of these studies involve less dependence on test-retest stability as an appropriate criterion for the comparative evaluation of adaptive testing strategies in live-testing studies. Instead, they use parallel-forms reliabilities as an appropriate evaluative criterion. Studies using test-retest stability are carefully designed to equate testing strategies for number of items administered, item discriminations, and memory effects.

While the computer simulation studies have resulted in useful information concerning the relative performance characteristics of a number of adaptive testing strategies, the results to date are limited in their generality. This derives from the fact that most of the computer simulation studies have relied heavily on information curves as an evaluative criterion. However, in the later phases of the research program it became obvious that other criteria, such as bias and the nature of the regression of ability estimates on true ability, are appropriate characteristics of adaptive testing strategies to be studied. Also, the computer simulations directly comparing adaptive testing strategies which have been implemented involved a restricted set of assumptions. It will be necessary to evaluate various adaptive testing strategies in terms of a variety of evaluative criteria, under a variety of conditions, with a carefully constructed set of item pool configurations, and at different test lengths. Such studies are currently in progress.

A further extension of the research effort would include evaluation of some adaptive testing strategies not studied to date. Specifically, the maximum likelihood adaptive testing strategies have not been studied in the present research, in comparison with other strategies. Furthermore, certain variations of some of the present adaptive testing strategies are also in need of research. For example, further research is necessary to develop termination criteria for use in the stradaptive and in the Bayesian testing strategies. Research is also necessary to develop

hybrid adaptive testing strategies which combine the desirable features of several of the approaches studied to date, and to study their psychometric characteristics.

Scoring methods. Research has shown that different scoring methods lead to ability estimates with different characteristics. Thus, the development of optimal scoring methods for adaptive testing is another important area for future research. Research on scoring methods should include the evaluation of the resulting ability estimates on a variety of criteria. In addition, the characteristics of these scoring methods should be studied in a variety of item pool configurations using a number of adaptive testing strategies.

In addition to evaluating the relative performance characteristics of a variety of scoring methods designed for dichotomous response testing, research should also proceed in the utilization of different modes of response in adaptive testing. In contrast to paper-and-pencil testing, computerized testing permits an immediate evaluation of the admissibility of each item response made by a testee. Consequently, it is possible to implement both graded and continuous methods of responding to ability tests within the framework of computerized adaptive testing. The use of these methods will result in new branching models and scoring methods, and considerable research will be necessary to determine the utility and adequacy of the non-dichotomous response modes possible in computerized ability testing. Furthermore, the use of these methods of responding to ability tests, as well as free-response items, might result in a reduction of guessing behavior, which was found to seriously affect the performance characteristics of some adaptive testing strategies. Similarly, research with non-dichotomous methods of response might lead to the development of methods to determine when a testee is guessing in response to a given test item.

Dimensionality. The finding that individuals differ in their response consistency in adaptive tests, and that response consistency appears to be a moderator of test-retest stability, implies that additional research needs to be done in the area of the dimensionality of individual response records. If high consistency in a stradaptive test is interpreted as unidimensionality of the response record, this suggests that low consistency reflects non-unidimensionality. Consequently, it should be possible to develop measures of an individual's response dimensionality during the process of adaptive testing. When the computer recognizes that an individual is responding in a multidimensional fashion, as opposed to the usually assumed unidimensional model, it could be programmed to implement multidimensional adaptive branching strategies for that individual. This, of course, implies that it will be necessary to develop and refine multidimensional adaptive testing strategies to correct for intra-individual variations in dimensionality.

The fact that many abilities are correlated to varying degrees suggests that the optimal implementation of computerized adaptive testing would take into account the intercorrelations among ability

dimensions. Consequently, research is needed on the development of branching models which account for inter-variable multidimensionality. One approach to this problem which will be pursued during the next phases of this research program is the development of branching techniques to locate an individual's position in a continuous multidimensional ability space.

Psychological effects. Studies of the psychological effects of adaptive testing and immediate knowledge of results suggest that both are important variables for further research. A fertile area for investigation is that of the difficulty of a test which will result in optimal levels of motivation and performance for individuals. Also important is the role of immediate knowledge of results in ability testing. One central question is whether feedback should be systematized as computer-administered feedback or should occur as self-administered feedback based on each testee's ability to infer the correctness of each item response. Thus, this line of research should result in the development of an adaptive test which administers to an individual the items that will result in optimal motivation and performance, as a consequence of the proportion of positive feedback that the individual obtains from the testing experience.

Observations during the testing of thousands of subjects in this research program indicate that the characteristics of the computer system and the terminal displays may have differential psychological effects on testees. Consequently, it is important that research be done on the nature of the computer terminal which will provide least interference with a testee's ability-testing behavior. This should include study of the display speed of the terminal and the visual characteristics of the terminal display.

The fact that large-scale time-shared computer systems tend to have unpredictable response times implies the use of smaller scale, preferably real-time, computer systems for the implementation of adaptive testing. However, a good minicomputer system can result in extremely fast response times (less than one-quarter second) following a testee's answer to a given test question. Thus, an important question for research is whether such very fast response times will result in a testee feeling unnecessarily "paced", with resulting increases in detrimental test anxiety and decreases in test-taking motivation. Research is necessary to determine the optimal computer response time, for the average testee and for specific testees. Similarly, the display speed of the terminal is another important factor which may differentially affect testee behavior. Display speeds of 10 characters per second are much too slow for the average testee, and display speeds of 960 characters per second may be too fast. Research should be designed to study the effects of different display speeds on testee behavior, holding constant other characteristics of the testing environment.

New tests. The majority of research to date in computerized adaptive testing has been within the framework of inter-item branching models,

using multiple-choice tests similar to those used in conventional paper-and-pencil testing. However, these approaches do not fully utilize the capabilities of computer systems. Thus, research is necessary that will develop computerized ability tests which take into account the unique capabilities of interactive computer equipment. Rather than using simple item-to-item branching techniques, this research would reconceptualize ability into an interactive problem-solving environment. The testee would be presented with a problem, and he/she would interact with the computer to attempt to solve the problem. The development of such new methods of testing will require new methods for evaluating an individual's interaction with the computer during the solution of a specified problem. These modes of interaction would include the individual's method of solving the problem, an evaluation of the quality of the final solution, and measurements of the speed with which the final solution was reached. Clearly, considerable new ground will need to be broken in this area of interactive testing.

Research is also needed on new forms of item presentation. This includes test items which are not in the typical multiple-choice format, and items which are pictorial in nature. For both kinds of items, testees should be able to respond in non-verbal ways where possible, or in natural language. Clearly, both computer hardware and software developments, as well as a generalization of psychometric theory, will be required to implement some of these new modes of testing.

Similarly, the capability of measuring abilities which are now poorly measured on paper-and-pencil tests would be a fertile area for further research in computerized adaptive testing. These include the measurement of memory abilities, the measurement of movement-based abilities, and the measurement of decision-making abilities. The inclusion of such tests in a computerized adaptive testing battery, possibly combined with existing tests based on a dimensional conceptualization of human abilities, could result in substantially greater validity in the prediction of practical criteria than is now possible with conventional paper-and-pencil ability tests.

ABSTRACTS OF RESEARCH REPORTS

Research Report 73-1
Ability Measurement: Conventional or Adaptive?
David J. Weiss and Nancy E. Betz
February 1973

Research to date on adaptive (sequential, branched, individualized, tailored, programmed, or response-contingent) ability testing is reviewed and summarized, following a brief review of problems inherent in conventional individual and group approaches to ability measurement. Research reviewed includes empirical, simulation, and theoretical studies of adaptive testing strategies. Adaptive strategies identified in the literature include two-stage and multistage tests. Multistage tests are differentiated into fixed-branching models and variable-branching models (including Bayesian and non-Bayesian strategies). Results of research using the various strategies and research approaches are compared and summarized. These studies lead to the general conclusion that, under a number of circumstances, adaptive testing can considerably reduce testing time and at the same time yield scores of higher reliability and validity than conventional tests. A number of new psychometric problems raised by adaptive testing are discussed, as is the criterion problem in evaluating the utility of adaptive testing. Problems of implementing adaptive testing with paper and pencil or with special testing machines are reviewed; the potential advantages of computer-controlled adaptive test administration are described. (AD 757788)

Research Report 73-3*
The Stratified Adaptive Computerized Ability Test
David J. Weiss
September 1973

The stratified adaptive (stradaptive) test is described as a strategy for tailoring an ability test to individual differences in testee ability levels. Stradaptive test administration is controlled by a time-shared computer system. The rationale of the method is described, which derives from Binet's strategy of ability test administration and findings concerning peaked tests from modern test theory. The essential elements of stradaptive testing considered include the differential entry point, branching rules, and individualized termination criteria. Different methods of scoring the stradaptive test are discussed, as are the implications of individual differences in consistency of test responses within the stradaptive test record. Examples of the results of live stradaptive testing are presented and discussed. Implications of additional data derived from stradaptive test response records are considered and related to other psychometric concepts. (AD 768376)

---

*Research Report 73-2 was not supported by this contract.

Research Report 73-4

An Empirical Study of Computer-Administered Two-Stage Ability Testing

Nancy E. Betz and David J. Weiss

October 1973

A two-stage adaptive test and a conventional peaked test were constructed and administered on a time-shared computer system to students in undergraduate psychology courses. Comparison of the score distributions showed that the two-stage test scores were somewhat more variable than the conventional test scores. The comparison also showed that the distribution of the two-stage scores was normal, whereas that of the conventional test scores tended toward flatness. The two-stage test had higher test-retest stability than the conventional test when the effect of memory was considered. The relationship between the two-stage and conventional test scores was relatively high and primarily linear, but it left about 20% of the reliable variance in the conventional test scores unaccounted for. Further analyses of the two-stage test showed that the difficulty levels of the measurement tests were not optimal, and that 4 to 5% of the testees were misclassified into measurement tests. The relatively poor internal consistency of the measurement tests in comparison to those of the routing and conventional tests was apparently due to the extreme homogeneity of ability within the measurement test sub-groups. The findings of the study were interpreted as favorable to continued exploration of two-stage testing procedures. Suggestions for improving the characteristics of the two-stage testing strategy are offered. (AD 768993)

Research Report 74-1

A Computer Software System for Adaptive Ability Measurement

Louis J. DeWitt and David J. Weiss

January 1974

A system of computer programs designed to control the administration of adaptive ability tests was developed and used for over 2500 hours of ability measurement. The system is capable of administering any combination of two testing strategies to a given individual without interruption. Each test can be based on one of six different testing strategies and can administer items selected from up to nine different item pools within each strategy. The system is designed to accept either multiple-choice responses or free-response numeric responses. For each test and testee, the administrator can choose whether to give no feedback, feedback after each item, or item-by-item feedback upon completion of testing. The requirements of the research design and constraints of the computer system as well as practical considerations are detailed, and their role in the design of the system are discussed. Technical requirements of the software system and problems that might arise in a transfer of the software system to another computer system are considered. Some basic concepts of computer programming are developed as an aid to the reader not technically trained in computer concepts. (AD 773961)

Research Report 74-2

A Word Knowledge Item Pool for Adaptive Ability Measurement

James R. McBride and David J. Weiss

June 1974

A series of four vocabulary norming tests was used to develop a large homogeneous pool of vocabulary test items for use in computer-administered adaptive testing research. A total of 575 unique vocabulary knowledge items was divided among four norming tests and administered to separate groups of college undergraduates. Norming tests were administered by computer or paper-and-pencil in fixed and random order. Analyses showed no effects due to item order or mode of administration. Item difficulty and discrimination indices of both the classical test model and the normal ogive item model were derived on the norming data. On the basis of item analysis results, 369 items were selected as satisfactory for the adaptive testing item pool. Factor-analytic studies of subsets of the 369 items confirmed the assumption of unidimensionality of the selected item pool. On the basis of known technical limitations in the research and the unique problems of developing item pools for adaptive testing, an outline was developed for the design of future norming studies specifically intended to develop large homogeneous test-item pools for use in computer-administered adaptive ability measurement. (AD 781894)

Research Report 74-3

An Empirical Investigation of Computered-Administered Pyramidal Ability Testing

Kevin C. Larkin and David J. Weiss

July 1974

Three pyramidal adaptive tests and a conventional peaked test were constructed, and were administered by time-shared computer to two separate groups of undergraduate psychology students. Six different methods of scoring pyramidal tests were evaluated, with respect to score distributions, stability, the relationship among scoring methods, and the relationship between pyramidal scoring methods and scores on the conventional test. For both the pyramidal tests and the conventional test, score distributions were platykurtic and positively skewed. Two methods of scoring the pyramidal tests consistently used an equal or greater proportion of the range of possible scores than the conventional test. The 15-stage pyramidal tests showed test-retest correlations only slightly lower than those for the 40-item conventional test. However, when the effects of memory were considered, the pyramidal strategy yielded more stable ability estimates than conventional tests of equivalent length. The correlation between pyramidal and conventional test scores ranged from .82 to .86 depending on the scoring method used. One pair of scoring methods was found to be perfectly correlated for properly constructed pyramidal tests; a second pair correlated almost perfectly. Findings were generally in favor of pyramidal testing, but further investigation of this adaptive testing strategy is necessary to determine its other important psychometric characteristics and to develop optimal rules for constructing pyramidal item structures. (AD 783553)

Research Report 74-4
Simulation Studies of Two-Stage Ability Testing
Nancy E. Betz and David J. Weiss
October 1974

Monte carlo simulation procedures were used to study the psychometric characteristics of two two-stage adaptive tests and a conventional "peaked" ability test. Results showed that scores yielded by both two-stage tests better reflected the normal destribution of underlying ability. Ability estimates from one of the two-stage tests were more reliable and had a slightly higher relationship to underlying ability than did the conventional test scores. One of the two-stage tests yielded an approximately horizontal information function, indicating more constant precision of measurement for individuals at all ability levels. The conventional test and the second two-stage test yielded information functions peaked at the mean ability level but dropping off at more extreme levels of ability; however, the second two-stage test provided more information than the conventional test at all levels of ability. The findings of the study were interpreted as indicating the potential superiority of two-stage tests in comparison to conventional tests. Several improvements in the construction of two-stage tests are suggested for further research. (AD A001230)

Research Report 74-5
Strategies of Adaptive Ability Measurement
David J. Weiss
December 1974

A number of strategies are described for adapting ability test items to individual differences in ability levels of testees. Each strategy consists of a different set of rules for selecting the sequence of test items to be administered to a given testee. Advantages and disadvantages of each strategy are discussed, and research issues unique to the strategy are described. Strategies reviewed are differentiated into two-stage and multistage approaches. Several variations of the two-stage approach are described. Multistage strategies include fixed-branching and variable-branching strategies. Fixed-branching strategies reviewed include a number of variations of the pyramidal approach (e.g., constant step size pyramids, decreasing step size pyramids, truncated pyramids, multiple-item pyramids), the flexilevel test, and the stradaptive test. Variable-branching approaches include two Bayesian strategies and two maximum likelihood strategies. The various strategies are compared with each other on important characteristics and on practical considerations, and are ranked on their apparent potential for providing equally precise measurement at all ability levels. (AD A004270)

Research Report 75-1
An Empirical Comparison of Two-Stage and Pyramidal Adaptive Ability Testing
Kevin C. Larkin and David J. Weiss
February 1975

A 15-stage pyramidal test and a 40-item two-stage test were constructed and administered by computer to 111 college undergraduates. The two-stage

test was found to utilize a smaller proportion of its potential score range than the pyramidal test. Score distributions for both tests were positively skewed but not significantly different from the normal distribution. The pyramidal test's score distributions tended to be platykurtic while the two-stage test's distribution tended to be leptokurtic. The assignment of subjects to measurement subtests in the two-stage test was more accurate than in a previous empirical investigation, since the misclassification rate was less than 1%. Comparison of scoring methods for the pyramidal strategy supported earlier findings that the average difficulty scoring methods were most useful. The correlations between scores on the two adaptive strategies ranged from r=.79 to .84. Both adaptive strategies appeared to adapt item difficulties to individual ability differences so as to reduce chance effects due to guessing. The pyramidal strategy seemed to be slightly more successful in eliminating guessing than the two-stage strategy. Results are discussed with respect to internal consistency reliabilities, stabilities, and the relation of each strategy to conventional testing. Simulation studies are suggested to further delineate the optimum characteristics of each testing strategy. (AD A0667733)

Research Report 75-2
TETREST: A FORTRAN IV Program for Calculating Tetrachoric Correlations
James R. McBride and David J. Weiss
February 1975

A general purpose computer program for the calculation of a matrix of tetrachoric correlations is described. This program was developed for use in adaptive (and other) testing research to examine the unidimensionality assumption in latent trait theory, in conjunction with available factor analysis programs. Several other potential applications and details for its use are described. The program accepts as input raw dichotomous data, reduced joint frequency data, or joint and marginal proportions, for up to 75 items. Output options include the tetrachoric correlation matrix, the matrix of phi coefficients, fourfold frequency tables for every item pair, a joint frequency matrix (which reduces all the information in the fourfold tables to a square matrix with order equal to the number of items), and a pair-by-pair listing of input proportions and output correlations that permits testing the program against published tables of the tetrachoric correlation. Variable input and output formatting makes the program convenient to use in conjunction with other analyses by packaged statistical programs. Examples of input and output are presented. A complete FORTRAN IV listing is included. (AD A007572)

Research Report 75-3
Empirical and Simulation Studies of Flexilevel Ability Testing
Nancy E. Betz and David J. Weiss
July 1975

A 40-item flexilevel test and a 40-item conventional test were compared, using data obtained through 1) computer-administration of the two tests to

three groups of college students, and 2) monte carlo simulation of test response patterns. Results indicated that the flexilevel score distribution better reflected the underlying normal distribution of ability, and that the flexilevel test had a higher parallel-forms reliability and a higher relationship to underlying ability level than did the conventional test. The overall test-retest stability of the two tests was equivalent, but there was evidence indicating that memory effects inflated the stability of the flexilevel test scores less than that of conventional test scores. The flexilevel test provided more accurate measurement at almost all ability levels, although its information function was similar in shape to that of the conventional test. However, the interpretation of differences in the level of information provided were confounded by differences in the average discriminating power of the items in the two tests. The flexilevel test also appeared to reduce random guessing behavior in comparison to the conventional test. (AD A013185)

Research Report 75-4

A Study of Computer-Administered Stradaptive Ability Testing

C. David Vale and David J. Weiss

October 1975

A conventional vocabulary test and two forms of a stradaptive vocabulary test were administered by a time-shared computer system to undergraduate college students. The two stradaptive tests differed in that one counted question-mark responses (i.e., omitted items) as incorrect and the other ignored items responded to with question marks. Stradaptive test scores were more consistent with the hypothesized nature of the population distribution of verbal ability. When corrected for differing levels of item discrimination and memory effects, the test-retest stabilities of the two testing strategies were about equal. Scores on one form of the stradaptive test were found to be very stable for testees with highly consistent response records on initial testing. Stability of "subject characteristic curve" data was high, suggesting the usefulness of these data for describing test-testee interactions. Of the ten stradaptive ability scores studied, which grouped into four clusters, average difficulty scores had the highest stabilities. Analysis of difficulties of items associated with correct, incorrect, and question-mark responses suggested that items with question mark responses should not be ignored, but should be treated as incorrect responses in branching decisions. Suggestions for future research on the stradaptive testing model are made. (AD A018758)

Research Report 75-5

Computerized Adaptive Trait Measurement: Problems and Prospects

Nancy E. Betz, James R. McBride, James B. Sympson and C. David Vale

with contributions by R. Darrell Bock and Robert L. Linn

Edited by David J. Weiss

November 1975

This report presents the proceedings of a symposium presented at the Annual Convention of the American Psychological Association, August 30, 1975.

The symposium consisted of four papers and the comments of two discussants.

1. C. David Vale. *Problem: Strategies of Branching through an Item Pool.*
   This paper describes a variety of strategies for adapting tests to the trait level of each individual on the basis of the testee's responses to previously administered items. Based on data from computer simulations, the various strategies are compared in terms of the levels and shapes of information curves they provide under one particular set of conditions. Limitations of the data presented are discussed.

2. James R. McBride. *Problem: Scoring Adaptive Tests.*
   Several approaches to scoring adaptive tests are described. Inapplicability of traditional number-correct scores in adaptive testing, where different individuals answer different items, is discussed. The essentials of latent trait theory are summarized, and two scoring methods usable with that approach are explicated. These scoring methods--maximum likelihood scoring and Bayesian scoring--are compared using simulation data, on criteria of information, bias, and regression on ability. Limitations of these scoring methods are discussed.

3. James B. Sympson. *Problem: Evaluating the Results of Adaptive Testing.*
   Six component elements of a testing procedure are described; it is suggested that proper evaluation of a testing procedure should be based on consideration of these elements as separable components. Classes of criteria for evaluating a testing procedure are differentiated into validating criteria, theoretical criteria, psycho-social criteria, and cost criteria. Within each of these categories, the various criteria are discussed and contrasted. Suggestions are made for the appropriate applications of each of these criteria. The problem of using multiple criteria is briefly discussed, and it is suggested that live-testing and simulation research be systematically combined. A number of specific recommendations are made concerning problems of evaluating the results of adaptive testing.

4. Nancy E. Betz. *Prospects: New Types of Information and Psychological Implications.*
   Several types of new information available from computerized adaptive measurement are described. These include individualized error of measurement, response consistency, improved response modes, response latencies, and new kinds of tests. Data from live computerized testing are presented showing that response consistency moderates test-retest reliability. The potential psychological advantages of computerized testing are discussed. Data are presented from two studies demonstrating the facilitating effect of immediate knowledge of results after each test item on ability test performance.

Comments by the discussants, Robert L. Linn of the University of Illinois and R. Darrell Bock of the University of Chicago, include a discussion of some of the limitations of the research presented, some differing interpretations, and suggestions for future research in adaptive testing. (AD A018675)

Research Report 75-6
A Simulation Study of Stradaptive Ability Testing
C. David Vale and David J. Weiss
December 1975

A conventional test and two forms of a stradaptive test were administered to thousands of simulated subjects by minicomputer. Characteristics of the three tests using several scoring techniques were investigated while varying the discriminating power of the items, the lengths of the tests, and the availability of prior information about the testee's ability level. The tests were evaluated in terms of their correlations with underlying ability, the amount of information they provided about ability, and the equiprecision of measurement they exhibited. Major findings were 1) scores on the conventional test correlated progressively less with ability as item discriminating power was increased beyond a=1.0; 2) the conventional test provided increasingly poorer equiprecision of measurement as items became more discriminating; 3) these undesirable characteristics were not characteristic of scores on the stradaptive test; 4) the stradaptive test provided higher score-ability correlations than the conventional test when item discriminations were high; 5) the stradaptive test provided more information and better equiprecision of measurement than the conventional test when test lengths and item discriminations were the same for the two strategies; 6) the use of valid prior ability estimates by stradaptive strategies resulted in scores which had better measurement characteristics than scores derived from a fixed entry point; 7) a Bayesian scoring technique implemented within the stradaptive testing strategy provided scores with good measurement characteristics; and 8) further research is necessary to develop improved flexible termination criteria for the stradaptive test. (AD A020961)

Research Report 76-1
Some Properties of a Bayesian Adaptive Ability Testing Strategy
James R. McBride and David J. Weiss
March 1976

Four monte carlo simulation studies were conducted of Owen's Bayesian sequential procedure for adaptive ability testing. Whereas previous simulation studies of this procedure have concentrated on evaluating it in terms of the correlation of its test scores with simulated ability in a normal population, these four studies explored a number of additional properties, both in a normally distributed population and in a distribution-free context. Study 1 replicated previous studies with finite item pools, but examined such properties as the bias of estimate, mean absolute error, and correlation of test length with ability. Studies 2 and 3 examined the same variables in a number of hypothetical infinite item pools, investigating

the effects of item discriminating power, guessing, and variable vs. fixed test length. Study 4 investigated some properties of the Bayesian test scores as latent trait estimators, under three different configurations (regressions of item discrimination on item difficulty) of item pools. The properties of interest included the regression of latent trait estimates on actual trait levels, the conditional bias of such estimates, the information curve of the trait estimates, and the relationship of test length to ability level. The results of these studies indicated that the ability estimates derived from the Bayesian testing strategy were highly correlated with ability level. However, the ability estimates were also highly correlated with number of items administered, were non-linearly biased, and provided measurements which were not of equal precision at all levels of ability.

Research Report 76-2
Effects of Time-Limits on Test-Taking Behavior
T.W. Miller and David J. Weiss
April 1976

Three related experimental studies analyzed rate and accuracy of test response under time-limit and no-time-limit conditions. Test instructions and multiple-choice vocabulary items were administered by computer. Student volunteers received monetary rewards under both testing conditions. In the first study college students were blocked into high- and low-ability groups on the basis of pretest scores. Results for both ability groups showed higher response rates under time-limit conditions than under no-time-limit conditions. There were no significant differences between time-limit and no-time-limit accuracy scores. Similar results were obtained in a second study in which each student received both time-limit and no-time-limit conditions. In a third study each testee received the same testing condition twice and higher response rates were observed under the time-limit condition; response accuracy remained consistent across testing conditions. All three studies showed essentially zero correlations between response rate and response accuracy. Response latency data were also analyzed in the three studies. These data suggested the existence of different test-taking styles and strategies under time-limit and no-time-limit testing conditions. The results of these studies suggest that number-correct scores from time-limit tests are a complex function of response rate, response accuracy, test-taking style and test-taking strategy, and therefore are not likely to be as valid or useful as number-correct scores from no-time-limit tests.

Research Report 76-3
Effects of Immediate Knowledge of Results and Adaptive Testing on Ability Test Performance
Nancy E. Betz and David J. Weiss
May 1976

This study investigated the effects of immediate knowledge of results (KR) concerning the correctness or incorrectness of each item response on a computer-administered test of verbal ability. The effects of KR

were examined on a 50-item conventional test and a stradaptive ability test and in high- and low-ability groups. The dependent variable was maximum likelihood ability estimates derived from the item responses. Results indicated that mean test scores for the high-ability group receiving KR were higher than for the no-KR group on both the conventional and stradaptive tests. For low-ability examinees, mean scores were higher under KR conditions than under no-KR conditions on both tests, but the difference was statistically significant only for the conventional test. However, the higher mean scores of the low-ability testees on the stradaptive test indicated that for low-ability examinees, adaptive testing had the same incentive effects as did the provision of immediate KR. Knowledge of results did not have significant effects on either response consistency on the stradaptive test or response latencies, and neither the shapes of the resulting test score distributions nor the internal consistency reliability of the conventional test differed consistently as a function of KR conditions. No significant score differences were found on a 44-item post-test administered without KR, indicating that the facilitative effects of knowledge of results on test performance were confined to the test in which KR was provided. The results of the study were interpreted as indicating the potential of both immediate knowledge of results and adaptive testing procedures to increase the extent to which ability tests measure the "maximum performance" capabilities of each individual.

Research Report 76-4
Psychological Effects of Immediate Knowledge of Results and Adaptive Ability Testing
Nancy E. Betz and David J. Weiss
May 1976

This study investigated the effects of providing immediate knowledge of results (KR) and adaptive testing on test anxiety and test-taking motivation. Also studied was the accuracy of student perceptions of their test performance on adaptive and conventional tests administered with or without immediate knowledge of results. Testees were 350 college students divided into high- and low-ability groups and randomly assigned to one of four test strategies by KR conditions. The ability level of examinees was found to be related to their reported levels of motivation and to differences in reported motivation under the different testing conditions. Low-ability examinees reported significantly higher levels of motivation on the stradaptive test than on the conventional test, while the reported motivation of high-ability examinees did not differ as a function of testing strategies. The effect of knowledge of results on reported motivation also differed as a function of ability level. Low-ability testees reported lower motivation under KR conditions than under no-KR, while higher ability testees reported higher motivation with KR. Analysis of the anxiety data indicated that students reported significantly higher levels of anxiety on the stradaptive test than on the conventional test. The provision of KR did not result in significant differences in reported anxiety. However, highest levels of anxiety were reported by the low-ability group on the stradaptive test administered with KR. These results, in conjunction with previously reported data on effects of KR on ability

test performance, were interpreted as being the result of facilitative anxiety. Over 90% of the students reacted favorably to the provision of immediate knowledge of results. Students were able to perceive their levels of test performance with some accuracy. However, perceptions of the relative degree of test difficulty were much more closely related to actual test score on the conventional test than on the stradaptive test. Thus, it appears that adaptive testing creates a psychological environment for testing which is more equivalently reinforcing or encouraging for examinees of all ability levels.